Panorama de l'information géologique armoricaine sur l'internet : méthodologie de recherche d'information, outils de veille automatisée

Alain-Hervé Le Gall, Isabelle Dubigeon

Panorama de l'information géologique armoricaine sur l'internet : méthodologie de recherche d'information, outils de veille automatisée

Alain-Hervé Le Gall [1] et Isabelle Dubigeon [2]

[1] *Géosciences Rennes – Centre commun de ressources et de documentation du CAREN, Université de Rennes 1, Bâtiment 14B, CS 74205, 35042 RENNES CEDEX. Tél : 02.23.23.60.75 Fax : 02.23.23.60.77*
Email : ahlegall@univ-rennes1.fr

[2] *Géosciences Rennes – Service Documentation Communication, Université de Rennes 1, Bâtiment 15, CS 74205, 35042 RENNES CEDEX. Tél : 02.23.23.58.10 Fax : 02.23.23.67.80.*
Email : isabelle.dubigeon@univ-rennes1.fr

RESUME : Rechercher de l'information sur l'internet… c'est facile ! Tout est sur le web, et puis les outils de recherches sont tellement intelligents et rapides ! Pas si simple, nous le savons bien : à la mi-octobre 2004, une recherche sur Google donne 6640 pages web contenant les mots « massif » et « armoricain »…; pour « geologie » et « bretagne », ce sont 17600 résultats qu'il faudrait consulter, évaluer, trier ! Le but de cet article est donc d'abord et avant tout de proposer une méthodologie de recherche d'information sur l'internet, de rappeler la typologie (les caractéristiques) des outils pour rechercher cette information (annuaire, moteur, métamoteur, bases de données, agent client type Copernic, etc.), ainsi que les outils (comme WebSite-Watcher) indispensables pour surveiller les contenus (veille informationnelle automatisée sur les mises à jour de sites). Les mots-clés utilisés pour illustrer l'utilisation de ces outils concernent bien entendu la géologie du Massif armoricain. En fin d'article, nous proposons un lien vers une présentation normalisée, homogène, de ressources internet - représentative mais non exhaustive -, sur la géologie du massif armoricain : parties de sites web, sites web entièrement dédiés, bases de données, etc., représentant aussi bien des sites institutionnels, pédagogiques, qu'associatifs ou personnels.

MOTS-CLÉS : Massif armoricain, géologie de la Bretagne, sites web, outils de recherches, information, méthode de recherche

ABSTRACT : Searching for information on the internet… so easy isn't it ? All the informations we are searching for obviously exist on the net, and web tools are so intelligent and so fast ! But we all know that it is not so simple : at mid-october 2004, 6640 web pages contain both the words such as « massif » and « armoricain »… ; for « geologie » and « bretagne », it will be nearly 17600 resultas that we should analyze, evaluate, sort and so on ! The main purpose of this article is to propose a methodogy for searching information on the net, and to present the typology (the most important characteristics) of the internet tools for searching for these informations (directory, search engine, metacrawler, database, software agent such as Copernic, etc.) ; even

to present some tools like WebSite-Watcher that allow to monitor selected websites for updates and changes. The keywords taken to illustrate our search examples concern the geology of the armorican massif. At the end of the article, we propose a link to a standard presentation of websites (a representative but not exhaustive list) relating to the geology of the armorican massif : parts of websites, websites dedicated to this topic, databases, etc., concerning institutional, academic, society or personnal websites.

KEYWORDS : Armorican massif, geology of Brittany, websites, web tools, information, search methodology

1 - Méthodologie de recherche d'information

1.1 - L'internet : les caractéristiques générales

Le volume du web

Difficile à évaluer... du fait de l'absence d'inventaire systématique des sites, et des limites techniques des meilleurs moteurs de recherche. On estime à :

- 21 millions de noms de domaines dans le monde avec une terminaison de pays (.fr par exemple) selon le RIPE Directoy (source : *http://www.ripe.net/ripencc/pub-services/stats/hostcount*) sur 65 millions au total selon Vnunet (source : *http://www.vnunet.fr/cpres/art.htm?id=4006&date=2004-10-05*). La différence est constituée par les .com, les .org, etc. Soit une croissance annuelle d'environ 15% ; soit 6 milliards de pages (source: *http://www.cyveillance.com/web/newsroom/press_res.htm*) ;

- 84 milliards de pages avec le web invisible (pages générées suite à l'interrogation d'une ressource par un formulaire :
catalogues, bases de données, etc.)

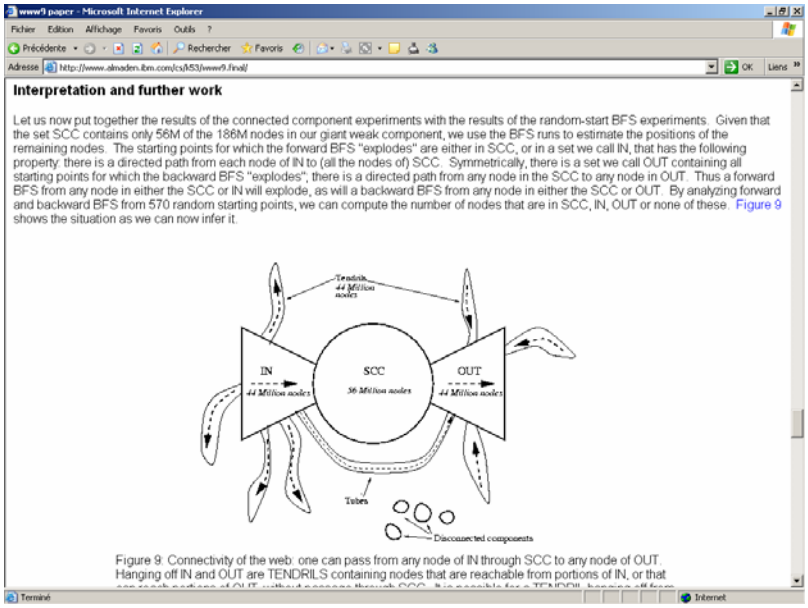
(source*http://brightplanet.com/deepcontent/deep_web_faq.asp#InvisibleWeb*) ;
- le volume d'ajout de pages uniques par jour était de 7,3 millions en 2001 (source: *http://c.asselin.free.fr/french/webenchiffre.htm#taille*).

La forme du web

Quelle est en quelque sorte. la topologie du web ? Deux théories se sont pendant quelque temps affrontées : « toile d'araignée » ou « noeud papillon » ?

Pendant longtemps, on a pensé que l'image du web visible s'apparentait à une toile d'araignée : avec des pages bien connectées entre elles. On imaginait donc qu'un moteur de recherches pouvait théoriquement tout visiter et tout indexer. C'était la théorie des « 19 clics » : deux pages quelconques du web étaient joignables au maximum après 19 liens successifs.

Mais en juin 2000, une étude menée par IBM, Compaq et Alta Vista a montré que la structure macroscopique du web était beaucoup plus complexe : avec une représentation graphique dite en "noeud papillon" :



Graph structure in the web. Andrei Broder, Ravi Kumar, Farzin Maghoul et al. (http://www.almaden.ibm.com/cs/k53/www9.final/)

Cette étude montrait l'existence de différentes parties rendant la navigation sur le Web difficile, voire impraticable, du fait de l'absence de liens hypertextes entre elles.

L'analyse de 200 millions de pages Web a montré que le Web était divisé en quatre grandes zones :

- le "noeud", qui est constitué du "noyau ultra-connecté" : celui-ci contient environ 56 millions de pages indexées en priorité par les moteurs (soit 28 % des pages) ;

- la partie gauche du "noeud papillon" représente 21 % du réseau (44 millions de pages) : ces pages permettent l'accès au coeur du Web mais l'inverse n'est pas possible. Cette partie du web est constituée de pages d'intérêt secondaire (pages personnelles), de pages récentes (pas encore signalées et liées au reste du web) ;

- la partie droite du "noeud papillon" représente 21 % du réseau (44 millions de pages) : c'est la logique inverse de l'aile gauche, les pages de destination sont accessibles depuis le coeur mais aucun retour n'est possible. Cette partie est constituée de sites commerciaux (sites vers lesquels pointent de nombreux liens mais qui, eux, n'en offrent que très rarement vers l'extérieur);

- une dernière zone est composée de pages non connectées au noeud : ce sont les « tendrils » (environ 20 % du web) : ce sont des pages accessibles depuis les pages d'origines et/ou qui donnent accès aux pages de destination;

- les pages restantes (soit 10 %) : elles sont totalement déconnectées du réseau (non liées). Il est donc impossible d'y accéder à moins d'en connaître préalablement l'adresse URL (http) exacte. De ce fait, elles échappent à l'indexation des moteurs de recherches.

L'organisation de l'information

Globalement, il n'y a pas ou peu d'organisation du contenu informationnel de l'internet : pas d'inventaire systématique, et donc pas de classement rationalisé. Les quelques tentatives d'organisation sont souvent désordonnées, non coordonnées : ce sont les annuaires de type yahoo!. L'internet n'est organisé qu'au niveau technique, autrement dit du point de vue du fonctionnement informatique du réseau (en somme la « tuyauterie »).

D'où la nécessité de s'en remettre à des outils de recherches : les annuaires, mais aussi les moteurs de recherches, les méta-moteurs... eux-mêmes pléthoriques.

Alors, comment chercher ?

1.2 ***– La recherche d'information : les principes généraux***

Une ébauche de méthodologie

Quelques conseils de bon sens :

- il faut « penser » sa recherche : préparer sa question, cerner sa problématique (définition du sujet, étude des concepts, définition d'une stratégie de recherche) ;

- choisir les bons mots-clés (voir ci-dessous) ;

- choisir et maîtriser les bons outils de recherche : les annuaires, les moteurs et métamoteurs ; et surtout jouer la complémentarité entre les différents outils, et ne pas se focaliser sur un seul d'entre eux ;

- conjuguer habilement ses recherches : dans les ressources classiques (les pages statiques), le web invisible (bases de données, catalogues, etc.), la presse et les actualités (newsletters, etc.) ;

- ne pas perdre le fil : suivre ses objectifs et sa stratégie initiale (attention à l'hypertexte, donc l'hyperchoix...) ; la recherche sur le web est un processus itératif qui impose le passage par plusieurs modes d'accès à l'information ;

- être « agile » : lire en travers, lancer plusieurs recherches simultanées, rebondir d'une information à l'autre, d'un outil à l'autre ;

- se limiter dans le temps : l'information recherchée n'existe peut-être pas sur le web...;

- avec le temps, se donner des points de repères (des points de départ) : des « bons sites » que l'on organisera en signets (dans son carnet d'adresses, les favoris).

Choisir les bons critères de recherche

Il va sans dire qu'il faut bien choisir ses mots-clés :

- un ou plusieurs ? Il faut procéder progressivement en affinant : 1 ou 2 critères au départ comme l'expression exacte "massif armoricain", ou l'équation « massif AND armor* », ou « bret* AND geolog* » (même si dans ce cas l'équation n'est pas aussi rigoureuse) ;

- si l'on trouve plus de 100 résultats, on aura alors intérêt à ajouter un critère supplémentaire.

Attention à l'utilisation des critères discriminants (opérateur SAUF) : à ne pas utiliser d'emblée, la méthode peut faire passer à côté de pages pertinentes (notamment des pages de synthèses).

Attention également à l'utilisation des majuscules et des minuscules, et à l'accentuation : le minuscule non accentué est universel (l'influence anglo-saxonne...). Les moteurs de recherches, AltaVista par exemple, sont sensibles aux majuscules et aux signes diacritiques (contrairement à Google).

Et la troncature ? Sur certains outils, la troncature (à droite) est implicite (c'est le cas dans Google). Sur d'autres outils, la chaîne de caractères exacte est recherchée (Yahoo! par exemple) : le symbole le plus communément utilisé est alors l'étoile (*).

Penser bien sûr aux synonymes : pour optimiser une recherche, puisque les outils ne le font pas eux-mêmes (il n'existe pas encore de dictionnaires de synonymes implicites dans les outils de recherches), on aura avantage à utiliser l'opérateur OU entre ces différents critères.

Choisir les bons types d'outils de recherche

Ce qui est différent de choisir les bons outils de recherche...

Quel type d'outil : annuaire ou moteur de recherche ? Cela va dépendre de sa maîtrise de la recherche d'information sur le web et de sa connaissance du domaine d'investigation.

Disons pour faire court :

- en fonction du type de recherches :

* pour une recherche large, une première approche : on choisira les annuaires généralistes

* pour une information ponctuelle : les moteurs généralistes

* pour des données factuelles (thématiques, sectorielles, disciplinaires...) : les annuaires et outils spécialisés

* pour une information récurrente sur un sujet : une identification via des métapages ou des annuaires spécialisés, puis une recherche par navigation ; complétée par l'utilisation des métamoteurs off-line (veille)

* pour des noms ou des chaînes de caractères : les métamoteurs

- en fonction de sa connaissance du sujet :

* avec une faible connaissance du sujet : on choisira plutôt les annuaires pour repérer les bons sites et les bons mots-clés ; puis, la navigation sur les sites de références ; puis, une recherche sur les moteurs et les métamoteurs

* avec bonne connaissance du sujet : une navigation directe sur les sites de références, complétée par l'utilisation de moteurs et métamoteurs

C'est ce que nous allons faire maintenant.

1.3 – La recherche dans les annuaires de recherches
Les caractéristiques générales

Qu'est-ce qu'un annuaire ? C'est une collection généraliste ou spécialisée de sites web classés par catégories organisées hiérarchiquement (200000 catégories pour Looksmart *http://www.looksmart.com ;* 590000 pour l'Open Directory *http://dmoz.org ;* 900 pour Nomade *http://www.nomade.tiscali.fr/*).

La sélection et le classement dans les annuaires sont manuels (c'est un travail humain). Autrement dit, la présence d'un site est aléatoire : sous réserve de sa soumission par le webmestre, sous réserve de son intégration par les cyber-documentalistes (Nomade par exemple rejette 40 % des 600 demandes quotidiennes).

Quelques chiffres : autour de 2,5 millions de sites sont référencés par les grands annuaires mondiaux (sur les 36 millions au total...) ; environ 100 000 pour les annuaires francophones (+ 600 par jour pour Nomade) (Source : *http://searchenginewatch.com/reports/article.php/2156411).*
Quelles conclusions en tirer ? Les annuaires constituent des outils de première approche : ils permettent d'avoir une vue d'ensemble d'un domaine (les sites « incontournables »).

Quels modes de recherche utiliser dans les annuaires ?

- méthode 1, de façon hiérarchique à travers les catégories, mais c'est une méthode relativement aléatoire : il n'existe pas de normes de présentation de l'arborescence des rubriques (nombre, hiérarchie etc.) communes aux annuaires ; pas plus qu'il n'existe de normes de nommage des rubriques
- méthode 2, par mots-clés via le formulaire : la recherche se fait alors sur l'intitulé des rubriques, c'est à dire le titre du site, le résumé descriptif du site, les URLs ; il n'y a pas de possibilité de requêtes complexes et pas d'indexation des pages en texte intégral. En tout état de cause, la recherche par formulaire se fait prioritairement sur le catalogue des sites référencés, et pas sur l'ensemble du web (sauf si un annuaire s'associe à un moteur pour permettre à ses usagers d'obtenir une réponse à sa recherche).

Quelques exemples de recherches dans yahoo! : on tirera avantage à utiliser le formulaire de recherche avancée du guide web (*http://fr.search.yahoo.com/search/fr/options/index.html?p=geologie+bretagne&fr=&h=c*).

Cela donne :

- aucune réponse pour « geologie bretonne », « geologie armoricaine », « SGMB »

- 1 réponse pour « massif armoricain » : il s'agit du site Armorique Minéraux (*http://perso.wanadoo.fr/armorique.mineraux/AccueilSuite.html*)

Dans Nomade, le résultat des recherches est également négatif : il existe pourtant une rubrique « Géologie locale » !

(*http://www.nomade.tiscali.fr/cat/nature_sciences/sciences_la_terre/geologie_mineralogi/geologie_locale/*).

Les résultats des recherches dans les annuaires généralistes sont souvent décevants. D'une façon générale, un travail de référencement de nos sites web est à faire : yahoo! propose une inscription gratuite en ligne (*http://add.europe.yahoo.com/bin/add?2100064866+FR).*

Comment perfectionner la recherche dans les annuaires ? Comment trouver des annuaires d'outils de recherches spécialisés, autrement dit comment trouver des annuaires d'outils thématiques en géologie ?

Il existe sur le web des annuaires spécialisés dans le signalement de répertoires généralistes ou spécialisés, de moteurs généralistes ou spécialisés, de métamoteurs.

Quelques exemples d'annuaires d'outils mondiaux généralistes : FinderSeeker (*http://www.finderseeker.com)* ou Search Engine Guide (*http://www.searchengineguide.com),* pour les annuaires d'outils francophones : Indicateur (*http://www.indicateur.com),* ou Enfin (*http://www.enfin.com),* et pour les métapages thématiques, en fonction d'une discipline, Geo-Guide (*http://www.geo-guide.de).*
Le constat est là aussi décevant : il n'existe pas de sites référencés sur la géologie du massif armoricain.

L'utilisation des moteurs de recherches devient alors indispensable.

1.4 – Les moteurs de recherche

Les principes de fonctionnement

Un moteur de recherche est un outil automatique constitué de 3 éléments :

- un robot d'exploration (« spider » ou « crawler ») qui visite et rapatrie dans une base de données les millions (milliards) de pages web (champ « texte », « titre de la page » et « URL ») ;

- un index (automatique) qui conserve tous les mots-clés significatifs contenus dans les pages (+ les métadonnées si elles existent) ;

- un formulaire d'interrogation de l'index, i.e. un serveur web avec un formulaire qui permet la recherche dans l'index de la base de données.

La mise à jour de l'index est variable selon les moteurs : entre 2 et 6 semaines selon la performance de l'outil et/ou son amplitude de couverture du

web. A noter que les sites les plus visités et les plus souvent mis à jour sont revisités en priorité par les moteurs.

Quelques statistiques

A propos des principaux moteurs mondiaux :

- nombre de pages indexés par les moteurs (sept. 2003) : Google indexe 3,3 milliards de pages, AltaVista uniquement 1 milliard de pages (source : *http://www.searchenginewatch.com/reports/sizes.html)* ;

- nombre de recherches effectuées par jour : 250 millions sur Google, 18 millions sur AltaVista (source :

http://www.searchenginewatch.com/reports/perday.html).

Les langages de recherche : la syntaxe « standard »

Bien que chaque moteur ait son propre langage d'interrogation, des opérateurs apparaissent de façon récurrente :

- opérateur inclusif : +

- opérateur exclusif : -

- troncature à droite : ...*

- expression exacte : «...»

Mais les moteurs proposent maintenant systématiquement une interface de recherche guidée qui dispense de connaître le langage d'interrogation (ce sont les intitulés du type « More options », « Advanced search », « Power Search », etc., sur Google : http://www.google.fr/advanced_search?hl=fr).

Les modes d'utilisation

Par rapport aux annuaires, les moteurs rendent possibles les recherches complexes. Mais équations élaborées signifie-t-il automatiquement recherches précises ? Et donc résultats précis ?

On sait a priori que les moteurs ne sont pas exhaustifs : ils n'indexent qu'une fraction du web (3,3 milliards de pages pour Google sur 6 milliards de pages pour l'ensemble du web « statique », donc grosso modo la moitié).

Les algorithmes de pertinence (pour la présentation du classement des résultats des recherches), qui prennent en compte l'indice de popularité des sites, le fait ou pas que les organismes est payés leur référencement, et donc un bon classement…, ne permettent pas de corriger les limites d'une indexation « basique » en texte intégral. Par ailleurs, il n'existe pas ou peu d'accès au web invisible (les fameuses 100 000 bases de données).

On constate également que les moteurs sont peu performants sur l'information « fraîche » (au regard de la vitesse de renouvellement des indexes, i.e. 2 à 6 semaines) ce qui signifie par conséquent qu'une recherche en « temps réel » est parfaitement illusoire.

Les moteurs n'archivent pas les pages modifiées ou disparues. Il n'existe à notre connaissance qu'un site qui ait une prétention d'archivage de l'internet (*http://www.archive.org*).

En résumé, les moteurs sont certes de plus en plus puissant, mais le web grandit de façon exponentielle...

Quelques exemples de recherches dans un moteur généraliste comme google (http://www.google.fr/advanced_search) :

- pages contenant tous les mots :

. « geologie » et « armor » donne 1710 pages

. « massif » et « armor » donne 11400 pages

. « geologie » et « bretagne » donne 18400 pages

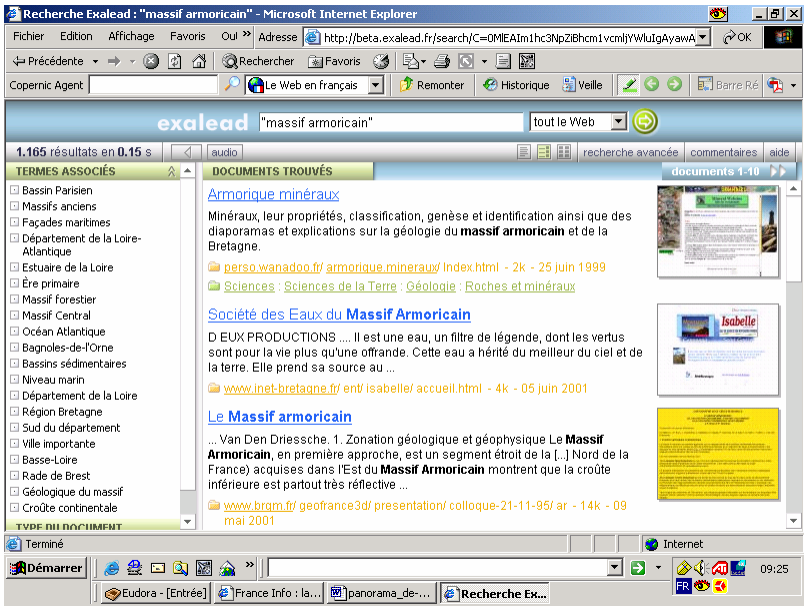
- pages contenant l'expression exacte :

. « massif armoricain » donne 5280 pages

. « geologie de la bretagne » donne 17 pages

. « geologie bretonne » donne 11 pages

Le moteur Exalead, un peu particulier, propose des mots-clés, en classant les résultats d'une recherche sous forme de sous-rubriques dans lesquelles il est possible de naviguer, ce qui peut être intéressant au début d'une recherche pour s'orienter dans les résultats et affiner la requête (http://www.exalead.fr/cgi/exalead/l=fr)



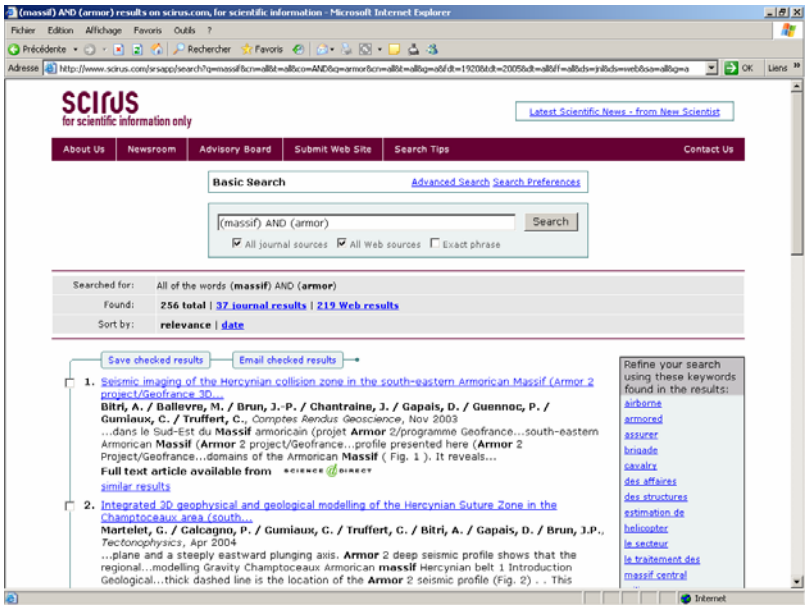
Les moteurs spécialisés

Ces moteurs sont dévolus à un domaine, une thématique, une discipline, etc. Il sont encore très peu nombreux. En réalité d'ailleurs, ces outils font une indexation en texte intégral des pages d'une sélection **manuelle** de sites.

Il en existe dans certaines disciplines scientifiques, citons pour l'exemple : en mathématiques et informatiques Fermivista (*http://fermivista.math.jussieu.fr*), en médecine (*http://www.mwsearch.com*) et en droit (*http://lawcrawler.findlaw.com*).

Il n'en existe pas en géologie, et encore moins sur le massif armoricain. On peut tenter sa chance sur des outils dédiés aux sciences en général, comme :

- Scirus, outil développé par l'éditeur Elsevier (*http://www.scirus.com*) : la recherche « massif » AND « armor » donne 256 pages dont 37 « journal results » (des articles dans des revues d'Elsevier) et 219 « Web results » (pages de sites).



Il faut tester également Scinet (*http://www.scinet.cc*) : la recherche « massif armor » donne 2505 résultats.

1.5 1.5 – Les métamoteurs « on-line »

Les principes de fonctionnement

Comment fonctionne un métamoteur « on-line » ? Dans un premier temps, le métamoteur interroge simultanément plusieurs moteurs et/ou annuaires classiques ; puis, dans un deuxième temps, il compile les résultats en faisant un dédoublonnage des résultats obtenus et en proposant un

nouveau classement. Autrement dit, un métamoteur ne gère pas de bases de données en propre : dans un premier temps, il fait travailler les autres outils de recherche.

Puis, dans un deuxième temps, il fait une concaténation des résultats obtenus.

Quels sont les avantages de ces outils ?

- ils sont efficaces et rapides pour des recherches de type « question-réponse » : i.e. les recherches simples ;

- ils donnent une idée du « répondant » des moteurs, et permet d'en choisir ensuite un plutôt qu'un autre ;

- ils pallient au manque d'exhaustivité inhérente à l'utilisation d'un seul outil.

Et les inconvénients ?

- ils sont incapables (pour le moment) de traduire systématiquement les différents langages d'interrogation : i.e. les syntaxes particulières à chaque outil ;

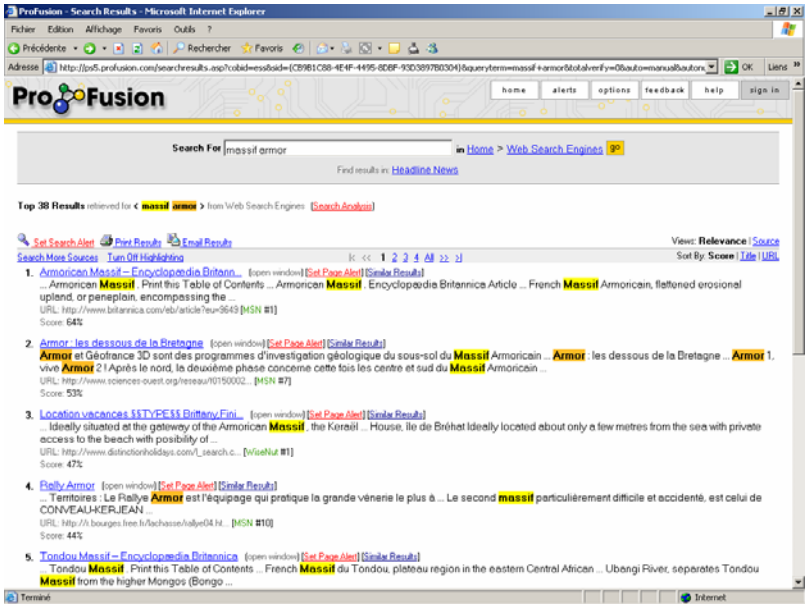
- les recherches complexes créent beaucoup de « bruit » ;

- ils ne sélectionnent souvent que les 10 premières réponses de chaque outil : or le classement des résultats d'une recherche est souvent… surprenant !

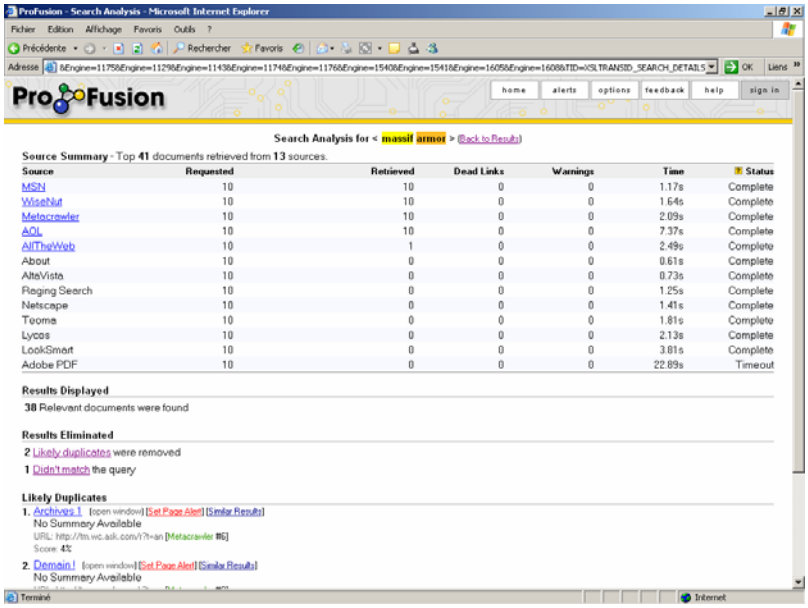
Quelques exemples de méta-moteurs

Les métamoteurs généralistes, les caractéristiques principales de chaque outil :

- Profusion (*http://www.profusion.com*) : la recherche « massif armor » s'affiche sous la forme suivante



Profusion fait également une analyse des résultats :



Source	Requested	Retrieved	Dead Links	Warnings	Time	Status
MSN	10	10	0	0	1.17s	Complete
WiseNut	10	10	0	0	1.64s	Complete
Metacrawler	10	10	0	0	2.09s	Complete
AOL	10	10	0	0	7.37s	Complete
AllTheWeb	10	1	0	0	2.49s	Complete
About	10	0	0	0	0.61s	Complete
AlteVista	10	0	0	0	0.73s	Complete
Raging Search	10	0	0	0	1.25s	Complete
Netscape	10	0	0	0	1.41s	Complete
Teoma	10	0	0	0	1.81s	Complete
Lycos	10	0	0	0	2.13s	Complete
LookSmart	10	0	0	0	3.81s	Complete
Adobe PDF	10	0	0	0	22.89s	Timeout

Search.com (*http://www.search.com*) : fait une recherche sur des outils spécialisés (Résultats à partir de : Google, LookSmart, MSN, Open Directory, Yahoo!, mySimon, etc.).

On peut également tester les outils suivants qui fonctionnent sur le même principe : Metacrawler (*http://www.metacrawler.com*) ou Ixquick

(*http://www.ixquick.com*) qui traduit les requêtes complexes à base de parenthèses.

Des métamoteurs spécialisés ont également vu le jour : ils interrogent simultanément sur le texte intégral de plusieurs bases de données thématiques comme en médecine Citeline (*http://www.citeline.com*) ou en droit Allaw (*http://www.allaw.com*).

Mais rien de très probant dans les autres disciplines scientifiques.

1.6 – Le web invisible

Définition

Le web invisible (ou « deep web ») est l'ensemble des pages non-localisables et/ou non indexables par les outils. A savoir :

- les documents dans des formats non HTML : pdf, ps, word, xls... (mais ce n'est plus vrai depuis 2001 avec Google notamment) ;

- les pages techniquement impossibles à indexer : frames, javascripts modifiant un contenu, flash, active X, java ;

- les pages isolées non reliées au reste de l'internet ;

- les pages accessibles à partir d'une authentification : mot de passe ;

- les pages paramétrées par une balise <robot> ou un fichier robot.txt exclusif (i.e. qui interdit sciemment aux moteurs d'indexer le contenu d'un site) ;

- les pages produites à partir de gestionnaire de site type ASP ou PHP : pages dynamiques produites « à la volée » ;

- les pages produites à partir de données rentrées dans un formulaire HTML : pages dynamiques générées par l'interrogation d'une base de données.

L'identification des ressources « invisibles »

Il y aurait environ 100 000 bases de données disponibles sur le web, soit potentiellement 84 milliards de pages dont 95 % accessibles gratuitement.

Comment les repérer ? Cela suppose a priori de posséder connaissance des ressources web dans sa discipline et de passer par les portails spécialisés.

Pour le repérage des bases de données, on peut utiliser des répertoires spécifiques comme :

- les bases de données gratuites sur l'internet (*http://urfist.univ-lyon1.fr/gratuits/index.html) ;*

- des banques de données pour les étudiants, les enseignants, les chercheurs (Formist) (*http://bdd.formist.enssib.fr/index.html) ;*

- les guides de recherches bibliographiques par disciplines (biblioguides de l'Université de Laval- Québec) (*http://www.bibl.ulaval.ca/info/biblgui.html) ;*

- les guides méthodologiques de recherches sur le web (par type de documents) comme SAPRISTI

(*http://docinsa.insa-lyon.fr/docinsa/sapristi/fristi31.html*.

Pour les ressources de ce type, ailleurs dans le monde, on utilisera :

- Invisibleweb (*http://www.invisibleweb.com*) qui recense 10 000 BdD

- ou CompletePlanet (http://www.completeplanet.com) qui recense 70000 BdD

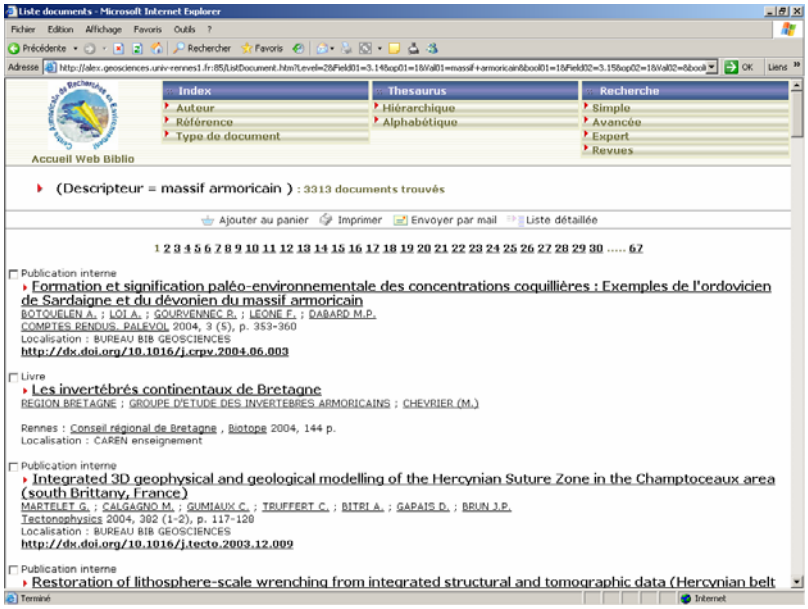
Aucune base de données spécialisée sur la géologie armoricaine n'est signalée.

Parmi les bases de données non-spécialisées mais potentiellement intéressantes, on trouvera :

- les catalogues de bibliothèques comme BN-OPALE, le catalogue de la BNF (*http://catalogue.bnf.fr/*) : la recherche « massif armoricain » dans le champ Sujet donne 33 réponses et 92 sur le champ Titre ;

- le SUDOC, le métacatalogue des bibliothèques universitaires françaises (*http://www.sudoc.abes.fr*) : la recherche « massif armoricain » donne 316 réponses (avec la localisation physique des documents dans leur bibliothèque ;

- Alexandrie, le catalogue de la bibliothèque du laboratoire de Géosciences Rennes et du CAREN (*http://alex.geosciences.univ-rennes1.fr:85/*) : la recherche « massif armoricain » dans le champ Descripteur donne 3313 réponses, dont 743 cartes, 643 thèses, 264 ouvrages, 1197 articles dépouillés (à noter qu'il nous reste environ 3000 tirés à part - de 1880 à nos jours - à enregistrer dans la base de données)



On peut également consulter les bases de données scientifiques multi-éditoriales qui sont en accès libre, comme :

- ArticlesSciences de l'INIST (http://articlesciences.inist.fr/) : la recherche « massif armoricain » n'importe où dans les notices (y compris le résumé) donne 75 réponses.

- la base de données PASCAL sur le serveur Biblioscience de l'INIST (accès réservé aux clients du CNRS) (http://www.inist.fr/bibliosciences/) : la recherche « massif AND armor* » n'importe où dans les notices (y compris le résumé) donne 1295 réponses.

- la base de données des Current Contents sur le serveur Biblioscience de l'INIST (accès réservé aux clients du CNRS) (http://www.inist.fr/bibliosciences/) : la recherche « massif AND armor* » n'importe où dans les notices (y compris le résumé) donne 108 réponses.

Les bases de données éditoriales sont également à consulter. L'accès aux références est gratuit, seul reste payante la consultation des articles :

- la plate-forme ScienceDirect de l'éditeur Elsevier (http://www.sciencedirect.com/) : la recherche « massif AND armor* » n'importe où dans les notices (y compris le résumé) donne 67 réponses.

Pour les autres plates-formes éditoriales, il faut penser à consulter également SpringerLink (http://springerlink.com/), Wiley Interscience (http://www3.interscience.wiley.com/), Kluwer Online (http://journals.kluweronline.com/), etc.

A noter que ces outils permettent de mémoriser un profil par mots-clés couplé à une alerte par mail, ce qui dispense à ressaisir manuellement et régulièrement son équation de recherche dans les diverses plates-formes.

1.7 1.7 – L'évaluation de l'information

Les critères d'évaluation

Une fois les sites web repérés, il s'agit d'évaluer la qualité de l'information qui y est proposée. Différentes catégories de critères sont à considérer :

- la pertinence : c'est à dire l'adéquation avec le sujet de la recherche et/ou de la veille ;

- la crédibilité : l'identification de l'organisme éditeur du site (organisme de recherches, association, sites personnel ?), des auteurs de documents, des cibles et objectifs du site ;

- la fraîcheur : la présence ou pas de la date de création et/ou de mise à jour des pages ;

- l'exactitude, l'exhaustivité : la qualité du contenu, la citation des sources, la présence d'une bibliographie, la qualité de l'expression, etc.;

- l'ergonomie : c'est à dire l'accessibilité au site, le temps de chargement des pages (notamment de la page d'accueil), qualité du système de navigation

(plan du site, moteur de recherche interne, contact avec le webmestre et/ou des personnes ressources), la qualité des applications, des base de données et programmes complémentaires utilisés (pdf, java, javascript, Flash, etc) ;

- le design : la présentation visuelle (l'organisation de l'information dans l'espace) et le graphisme.

Pour en savoir encore plus sur l'auteur d'un site

Pour trouver le propriétaire d'un nom de domaine, on peut consulter la base de données whois (http://www.andco.fr/).

Pour trouver des informations générales sur une page, l'utilitaire gratuit Alexa (http://info.alexa.com) permet d'obtenir : les coordonnées du propriétaire, les statistiques sur le site, le temps de chargement de la page, le nombre de liens vers le site ; des sites/pages similaires au site en cours de consultation ; en cas d'erreur 404 (« document not found ») : la copie de la page si elle est archivée par Alexa.

2 – Les outils de veille informationnelle

2.1 – Une typologie des outils

Définition

Les outils de veille sont des agents « évolués » plutôt que des agents « intelligents » : il y a pas ou peu d'intelligence artificielle, pas ou peu de technologies statistiques, etc.

Ils se présentent sous forme de logiciels à installer sur son disque dur. La version de base est souvent gratuite. Il sont essentiellement destinés à automatiser des tâches récurrentes comme :

- faciliter la navigation par une meilleure gestion de l'historique, du cache, des bookmarks (favoris), des informations sur les pages visitées ;

- aider à la recherche d'information : ces outils sont des métamoteurs relativement évolués et permettent une analyse linguistique des requêtes ;

- aider à l'exploitation des résultats : par des fonctions de tri, indexation, résumé automatique, export des résultats etc.;

- permettre une veille dans le temps : conserver les requêtes, archiver les résultats, mettre en exergue les modifications (trois actions qui ne sont pas possibles avec les recherches en ligne) ;

- permettre une diffusion sélective et personnalisée de l'information par mail.

Attention, ces outils sont « évolués » et non pas « intelligents » : l'intelligence humaine (connaissance du domaine, capacité de synthèse) reste indispensable, et il va de soi qu'un re-traitement de l'information est nécessaire. Plusieurs types d'outils de veille sont à notre disposition.

Les aspirateurs de sites

Ils se présentent sous forme de logiciels installés sur l'ordinateur, souvent gratuit dans leur version de base. Leur principe de fonctionnement est le suivant : ils procèdent à l'enregistrement du site sur le disque dur et utilisent une fonction de veille avec visite automatique des sites programmable et repérage des modifications depuis la dernière visite.

Memoweb (*http://www.goto.fr*) est un produit avec une interface en français :

- il charge tout ou partie d'un site web sur votre ordinateur ;

- avec un facteur de compression : il permet d'économiser jusqu'à 70% d'espace disque ;

- il constitue votre bibliothèque de documents internet : il permet de trier les ressources d'un site selon le type de document (page HTML, PDF, PPT, etc) et il extrait et exporte les adresses e-mail.

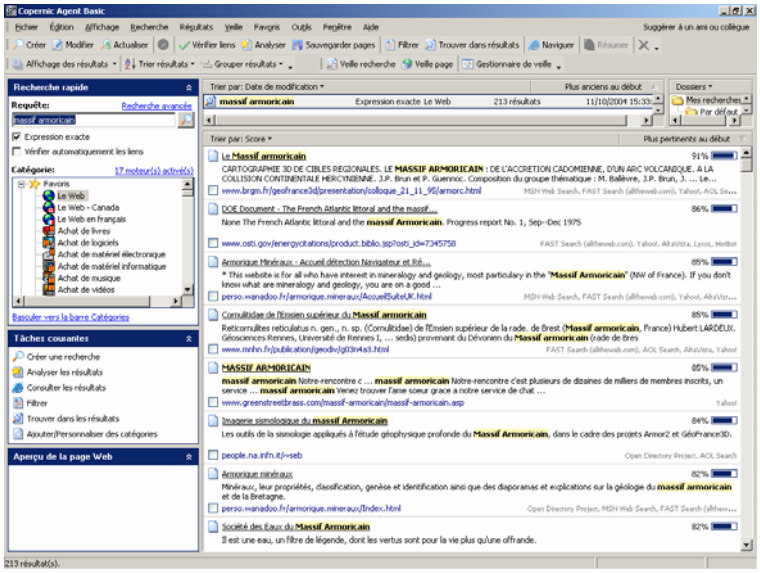
Les méta-moteurs « off-line »

Ce sont donc des métamoteurs clients : des logiciels installés sur l'ordinateur souvent gratuit dans leur version de base.

Ils ont les mêmes fonctionnalités de base que les métamoteurs on-line, avec des plus : veille, alerte, résumé, téléchargement des résultats, export HTML etc.

Il existe de très nombreux utilitaires : BullsEyes, Dig-Out, Strategic finder, Umap, WebCompass, Web Seeker etc.

Le plus connu d'entr'eux est Copernic (*http://www.copernic.com*) : la recherche de l'expression exacte « massif armoricain » donne 213 réponses



Fiche technique du produit et fonctionnalités (version de base) :

- domaines d'investigation : plusieurs outils francophones et/ou internationaux ;

- 3 modes de recherche : avec tous les mots (opérateur ET), un des mots (opérateur OU), l'expression exacte (guillemets), possibilité de paramétrer les requêtes ;

- équations de recherches : archivage dans des dossiers et mise à jour (programmable avec la version payante), possibilité de tri par mots-clés, date de dernière mise à jour ou domaine de recherche, etc.;

- résultats : ils apparaissent dans une fenêtre distincte avec un surlignage des mots-clés, l'indication d'un degré de pertinence, l'élimination des doublons, la possibilité de trier par le titre du document, l'adresse http, le score de pertinence, le moteur qui a trouvé la ressource ;

- téléchargement possible des résultats avec la consultation hors ligne et le raffinage : utilisation des opérateurs ET, OU, SAUF, PRES et le parenthèsage ;

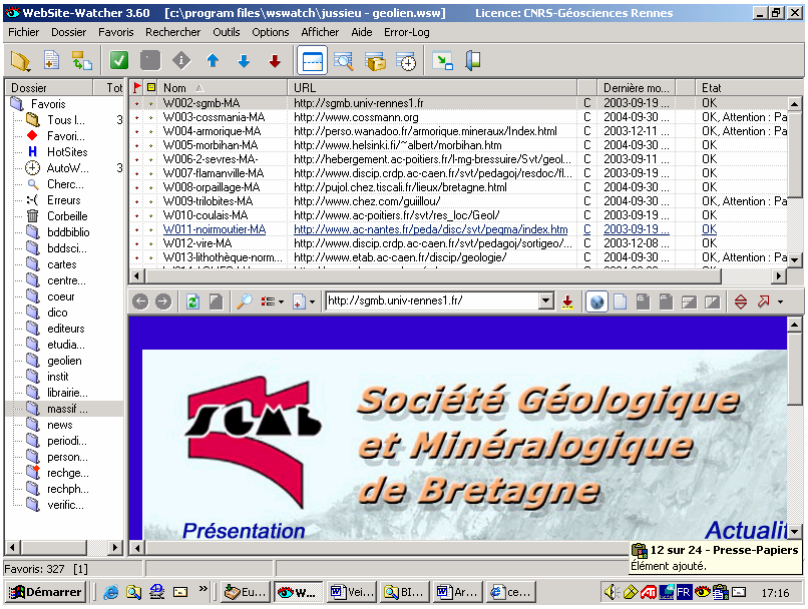
- la possibilité d'exporter en plusieurs formats TXT, HTML... et d'envoyer un résultat par mail aux formats TXT et HTML.

Les outils d'alerte

Ils signalent les mises à jour de pages web préalablement enregistrées *(« monotoring »).*

Ce sont des agents clients installés sur l'ordinateur, gratuit dans leur version de base

Website-Watcher (*http://aignes.com)* permet par exemple l'enregistrement des pages à surveiller, de définir le choix du système d'alerte (web et/ou mail), de spécifier certaines options de veille comme la modification de la page (surbrillance), l'archivage des pages modifiées "en local" (avec l'archivage des versions successives d'une page web).



D'autres outils du même type méritent d'être testés, comme Webspector (*http://www.illumix.com*) ou Web Track (*http://track.intelliseek.com*).

3.2 – Mise en place d'un système de veille

ATTENTION ! Ces outils simplifient certaines tâches fastidieuses, mais ils ne suppriment pas les risques inhérents à la veille :

- être noyé dans une masse de plus en plus importante d'information : l'information « web » est moins formalisée, moins structurée qu'une source traditionnelle, avec le problème crucial notamment de la fiabilité ;

- le rapport sur investissement n'est pas évident : le temps-coût / valeur de l'information est difficile à appréhender, alors que par ailleurs la veille est sensée faire gagner du temps...

En guise de conclusion, la méthodologie raisonnable à mettre en oeuvre est peut être la suivante :

- stocker les équations et les mots-clés ad hoc avec Copernic par exemple ;

- mémoriser les pages de sites avec Website-Watcher par exemple ;

- surveiller les sites dans leur ensemble avec MemoWeb par exemple.

La collecte de l'information se caractérise en général par une compilation d'informations fragmentaires, autrement dit de signaux « faibles » ; la sélection, quant à elle, consiste à affiner la collecte pour permettre l'analyse, et donc la création d'informations à valeur ajoutée.

Bibliographie :

Sur la recherche et la veille d'information sur l'internet

- livre : Foenix-Riou Béatrice. Recherche et veille sur le web visible et invisible. Edition tech & doc, 2001.

- en ligne :

ABONDANCE – Andrieu Olivier. Recherche d'information et référencement (*http://www.abondance.com*)

ADBS - Lardy Jean-Pierre. RISI : Recherche d'Information sur l'Internet (*http://www.adbs.fr/site/repertoires/sites/lardy/risi.htm*)

AgentLand, Site dédié aux outils de recherche et de veille sur l'internet (*http://www.agentland.f*)

CREPUQ (Conférence des recteurs et des principaux des universités du Québec). GIRI : Guide d'initiation à la recherche dans Internet (*http://www.bibl.ulaval.ca/vitrine/giri*)

GIRI2 : guide des indispensables de la recherche dans Internet (http://www.bibl.ulaval.ca/vitrine/giri/giri2/index.html)

INSA-Lyon. SAPRISTI : Sentiers d'Accès et Pistes de Recherche d'Informations Scientifiques et Techniques sur l'Internet (*http://csidoc.insa-lyon.fr/sapristi/digest.html*)

MEDIADIX (Centre de formation aux carrières des bibliothèques de l'Université de ParisX)

URFIST Paris – Tosello-Bancal Jean-Emile. Les outils de recherche d'information sur l'internet (*http://www.ccr.jussieu.fr/urfist/outsej.htm*)

Sur les ressources internet sur le Massif Armoricain

- en ligne :

Géosciences Rennes – Alain-Hervé Le Gall et Isabelle Dubigeon. Description normalisée de sites web sur le Massif armoricain (http://www.geosciences.univ-rennes1.fr/biblio/bibvirt/ressources/ressgeol/veille_MA_oct2004.PDF